Inter-Rater Reliability and Review of the VA Unresolved Narratives

J. Chris Eagon, MD, John F. Hurdle, MD, Ph.D., Michael J. Lincoln, MD
Department of Veteran's Affairs, IRMFO and Department of Medical Informatics, University of Utah

To better understand how VA clinicians use medical vocabulary in every day practice, we set out to characterize terms generated in the Problem List module of the VA's DHCP system that were not mapped to terms in the controlled-vocabulary lexicon of DHCP. When entered terms fail to match those in the lexicon, a note is sent to a central repository. When our study started, the volume in that repository had reached 16,783 terms. We wished to characterize the potential reasons why these terms failed to match terms in the lexicon. After examining two small samples of randomly selected terms, we used group consensus to develop a set of rating criteria and a rating form. To be sure that the results of multiple reviewers could be confidently compared, we analyzed the inter-rater agreement of our rating process. Two raters used this form to rate the same 400 terms. We found that modifiers and numeric data were common and consistent reasons for failure to match, while others such as use of synonyms and absence of the concept from the lexicon were common but less consistently selected.

INTRODUCTION

Few doubt the need to review and analyze clinical terms generated in medical practice. These terms are the best window we have into the nature of medical discourse. Our understanding of this discourse will determine how well we can employ computers in health care. This understanding of vocabulary and usage improves the quality of basic automation tasks (e.g., data recording, order entry, simple alerting), and paves the way for long-sought advanced clinical computing tasks (e.g., automated decision support [1-2] and natural language processing for querying, transcription, and text classification. [3-4]) In both cases, insight into vocabulary standards is essential.

We have initiated a bottom-up approach for analyzing the vocabulary used in VA's Decentralized Hospital Computer Program (DHCP). In 1994 the VA released a module for DHCP called "Problem List." Clinicians as well as clerical staff use it to specify patient problems. A set of tools (called Multiterm Lookup) and a clinical vocabulary (Lexicon Utility) were installed to assist in the automated mapping of free-text entries to standard concepts. Among other advantages, this lexicon can allow mapping of Problem List concepts to ICD-9CM, which will prove vital as the VA assumes more responsibility for patient billing.

The Lexicon Utility (LU) was seeded from a subset of the National Library of Medicine's 1992 Metathesaurus [5]. LU contains 65,718 major concepts and 96,246 total expressions. Problem List, which uses the LU, works by displaying a series of related LU concepts (in groups of five) whenever a user types in a free-text term. Once selected, an unambiguous mapping to a standard term fitting the patient's clinical presentation is recorded in the patient's record.

Problem List is used by physicians, nurses, and clerical staff to enter patient problems into the patient's computer-based medical record. Problems may subsequently be used for direct patient care, medical record review, resource accounting, and billing. It is possible for users to enter terms which are not mapped to LU terms, as will be described in the Methods section. How strictly the user adheres to utilizing terms from LU is site and user dependent. In cases where the user authorizes the use of the unmatched term, the free-text entry becomes the value stored as a problem for the patient and is called an "unresolved narrative" (UN). Because UNs are not mapped to a standard source, an important gap in the collection of coded data for research, clinical care, and billing occurs.

When an UN is created at a site, an entry is made in a local file noting the free text string, creation date and time, and the service of the user. After at least 50 UNs are collected at a site, an e-mail summary is sent to our VA software office for analysis. To assist in improving the Problem List module and the linkage of patient problems to coded terms, we wished to analyze why these terms were unresolved. The high volume of terms received (over 63,000 as of March 1996) requires multiple reviewers working in parallel. Relying on multiple reviewers raises the question of how well the different reviewers agree in their analysis. If we are to paint a consistent picture of the UNs, we need to assess the inter-rater reliability [6]. We report here our preliminary findings on this work.

METHODS

From April 1994 to August 1995, we received 16,873 UNs. The UNs were converted from the ASCII mailgrams sent by the various hospitals to a flat, field-delimited file, and these records were loaded into Microsoft Access, a relational database system. Fields in the main table included narrative text, creation date/time, VA site of origin, and service of the user who created the UN. Additional fields were later added to this table to record the rater's characterization of each narrative. The UNs were received from users assigned to 86 service units at 30 hospitals. The services of the users were grouped together into several broad categories: Medical Services-26%; Medical Admin.-19%; Research-14%; Ambulatory Care-14%; Medical Specialties and Clinics-8%; Nursing-8%; Other-11%.

Although we did not know the total number of problems entered by users, we did know the size of

0195—4210/96/$5.00 © 1996 AMIA, Inc.

the Problem List global variable at each VA site, and by estimating the average number of patient problems stored per 512 byte block of disc space (1.0-2.2), we estimated that UNs constitute 10-25% of the total number of problems.

We studied the 16,873 terms in two phases. The goal in each phase was to characterize the unresolved narrative along several dimensions: 1) the response of the Problem List module to the original free-text entry; 2) the degree to which the free text term matched a concept already in the LU; and 3) in cases where a related term *was* in the LU, the differences that had prevented the match between the UN and the related LU term; and 4) in cases where it did not match any LU term, the degree to which the term matched a term in the 1996 version of the NLM's Metathesaurus (Meta '96).

The Rating Procedure

In each part of the study, a random sample of the records was chosen for analysis. A sample of ten of these is shown in Table 1. Microsoft Access forms were designed that allowed each rater to see individual UNs and to select buttons indicating how that term was to be characterized without seeing the characterization of other raters. SQL queries were performed to characterize the ratings.

Raters used the Access form while running the Problem List application at the local VA Medical Center exactly as the user who created the term had done. The rater first verified that the user narrative was truly not contained in the LU. In most cases, there was not an exact match, so the rater would try different synonyms, alternate spellings, and fragments (stripped of modifiers) that looked to be a core concept in an attempt to characterize why the match did not succeed. The minimum alteration in the free text string required to generate a concept match was considered in categorizing the reasons for lack of match of the UN. By switching back and forth between the Problem List and Access, we could recreate the clinician's experience and record our characterizations simultaneously. One of the raters also had a window open to the current NLM Meta '96 Web site and identified UNs that matched Meta '96 using the Normalized String Index.

Phase I: Form Design and Small-Scale Trial

The University of Utah Department of Medical Informatics has performed several studies like the one we were undertaking, so we initially adopted a term ranking system and ranking form design based on their most recent effort [7]. After several iterations of design, we piloted a prototype form by having three reviewers (all MDs) independently rate the same 25 randomly-chosen terms. The primary focus of the rating was on the closeness of the match of the core concept, augmented by a characterization of the reason for the mismatch.

Table 1: Ten Random Sample Narratives

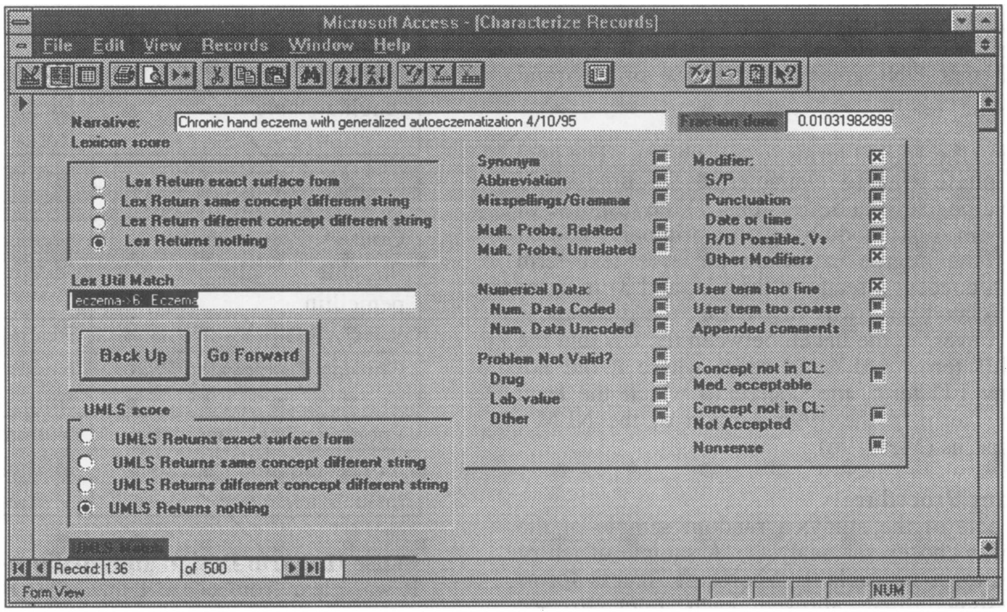
Narrative	Service of User
Angina-stable	Health Services R&D
414.00	Medical Admin.
Advice or health instruction	Medical Admin.
Gout vs. Pseudogout	Medical Service
Possible allergy to penicillin	Medical Service
S/P CABG	Extended Care
Chronically elevated PSA	Nursing Home Care Unit
Pneumococcal vaccination	Nursing
hx-prostatic malignancy	Medical Admin.
influenza vaccine	Medical Admin.

Phase II: Form Design and Medium-Scale Trial

It was clear from phase I that most narratives had a near or exact concept match and that more expansion of the lexical reasons for failure of MTLU to find the correct concept in LU was needed. The rating form was redesigned and is shown in Figure 1. The upper left-hand box captures what the LU returned when the exact surface form originally entered by the user is entered. LU can return, potentially, hundreds of terms in five member groupings. The response was placed into one of four mutually exclusive categories: Category 1- returns (in one of these groupings) the same string; Category 2- returns different strings but at least one with the same medical concept; Category 3- returns different strings and different concepts; or Category 4- returns nothing at all. The right-hand side of the form captures different reasons contributing to why the term did not match. These non-mutually exclusive reasons were clustered into groups and were selected if modulation of this aspect of the term was required for a close concept match. As an example of a synonym problem, the UN "bleeding gums" required substitution by "gingival hemorrhage" before it was matched by Problem List. Abbreviation was selected if expansion of an abbreviation in the UN was required to match the concept such as with "GER" to "Gastroesophageal Reflux." Categories were also included for when the raters were unable to identify a related concept in LU.

In this phase, two raters (both MDs) rated the same 400, randomly-chosen UNs, constituting 2.4% of the 16,873. Inter-rater agreement was measured using Cohen's kappa and Finn's r statistics, and bias between the raters was assessed using Bowker's extension of McNemar's test [8]. Cohen's kappa and McNemar's test were also used to measure agreement and bias on each reason for match failure using a Bonferroni correction for multiple (24) comparisons.

Figure 1: Sample Characterization Form



Narrative: Chronic hand eczema with generalized autoeczematization 4/10/95

Fraction done: 0.01031982899

Lexicon score

○ Lex Return exact surface form
○ Lex Return same concept different string
○ Lex Return different concept different string
● Lex Returns nothing

Lex Util Match: eczema->6: Eczema

UMLS score

○ UMLS Returns exact surface form
○ UMLS Returns same concept different string
○ UMLS Returns different concept different string
● UMLS Returns nothing

Synonym	▣	Modifier:	☒
Abbreviation	▣	S/P	▣
Misspellings/Grammar	▣	Punctuation	▣
Mult. Probs. Related	▣	Date or time	☒
Mult. Probs. Unrelated	▣	R/D Possible, Vs	▣
		Other Modifiers	☒
Numerical Data:	▣	User term too fine	☒
Num. Data Coded	▣	User term too coarse	▣
Num. Data Uncoded	▣	Appended comments	▣
Problem Not Valid?	▣		
Drug	▣	Concept not in CL: Med. acceptable	▣
Lab value	▣		
Other	▣	Concept not in CL: Not Accepted	▣
		Nonsense	▣

Record: 136 of 500

RESULTS

User-Interface Response Category Data

Figure 2 shows the response of Problem List Application to entry of the 400 UNs. Raters found the exact surface form of the UN returned in 3-4%, and a different term but the same concept returned in 7-10%. Only nonmatching concepts were returned in 3-5% and Problem List failed to return any potential match in 81-84% of UNs (95% CI: 77%-85%). By comparison, Metathesaurus returned exact or close concept matches for 17.2% of terms, returned only nonmatching concepts for 0%, and failed to identify any possible match in 82.8% of terms.

The raters were quite close most of the time. The coefficient of agreement *kappa* = 0.80 ($p < 0.0001$.) *Finn's r* was 0.76, indicating 76% of the inter-rater agreement was not due to chance. Bowker's extension of McNemar's test for symmetry of a 4x4 table showed a p-value > 0.10 indicating no systematic bias among the raters with regard to category selected. Even removing the large category 4 from the data, both raters gave similar results: a *kappa* of 0.68 with a $p < 0.0001$.

Reasons for Failure to Match

The raters selected one or more reasons for failure to match an exact string for all narratives in categories 2, 3, and 4. The groups of possible reasons for match failure are listed in Table 2, along with a breakdown of inter-rater agreement. Overall, the raters agreed on all selections of reasons for mismatch for 43.5% of UNs. In addition, on narratives where there was some disagreement, most selected reasons were still concordant for a given UN. The average number of reasons selected per UN is shown in the final row of table 2. The first column of data lists the number (and percentage) of terms, by reason, on which both raters selected that reason (they may have disagreed, though, on other reasons for a given term). The next column shows the number where at least one rater selected the reason. The last column shows the percentage of time that a reason selected by at least one rater was also selected by the other rater. The *kappa* statistic is also listed in parentheses. Only five low incidence reasons did not show significant agreement between the raters: Multiple unrelated problems, Drug, Punctuation, User term too coarse, and Medically unacceptable concept not in LU.

Based on McNemar's test, there was evidence for systematic bias among the raters in the reasons 'Synonym' and 'User term too fine.' A pattern seemed to emerge where one of the raters tended to call certain

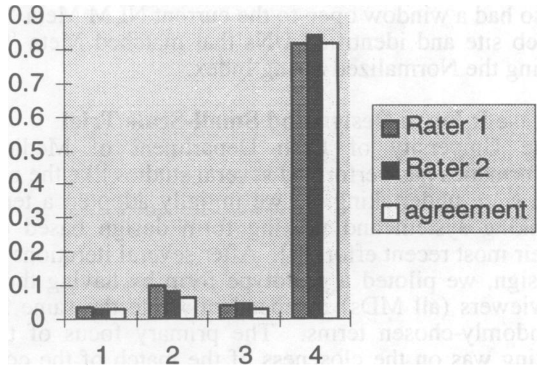


Figure 2: Fraction of Narratives by User Interface Response Category (1=Exact Match, 2=Diff. String & Same Concept, 3=Diff. String & Diff. Concept, 4=Nothing Returned)

Table 2: Reasons for Failure to Match LU Concept

	Category selected by both raters	Category selected by at least one rater	Fraction of agreement and *kappa*
N=400	n (percent)	n (percent)	percent (*kappa*)
Synonym	51 (13)	145 (36)	35 (.37)*#
Abbreviation	3 (1)	30 (8)	10 (.15)*
Misspelling/Grammar	16 (4)	37 (9)	43 (.58)*
Multiple Problems, Related	24 (6)	46 (12)	52 (.66)*
Multiple Problems, Unrelated	0 (0)	4 (1)	0 (0)
Numerical Data	44 (11)	50 (13)	88 (.93)*
Coded	42 (11)	47 (12)	89 (.94)*
Uncoded	2 (1)	3 (1)	67 (.80)*
Problem not valid	20 (5)	38 (10)	53 (.67)*
Drug	0 (0)	5 (1)	0 (0)
Lab Value	1 (0)	4 (1)	25 (.40)*
Other	17 (4)	31 (8)	55 (.69)*
Modifier	160 (40)	210 (53)	76 (.75)*
S/P	38 (10)	41 (10)	93 (.96)*
Punctuation	0 (0)	4 (1)	0 (0)
Date/time	5 (1)	13 (3)	38 (.55)*
R/O, Possible, etc.	5 (1)	10 (3)	50 (.66)*
Other	123 (31)	186 (47)	66 (.67)*
User term too fine	21 (5)	75 (19)	28 (.37)*#
User term too coarse	0 (0)	9 (2)	0 (-.01)
Appended comments	2 (1)	10 (3)	20 (.33)*
Medically acceptable concept not in LU	15 (4)	71 (18)	21 (.27)*
Medically unacceptable concept not in LU	0 (0)	4 (1)	0 (0)
Nonsense	2 (1)	9 (2)	22 (.36)*
Totals	591	1082	
Average number per UN	1.48	2.72	

*** *kappa* $p < 0.05$ # McNemar $p < 0.05$**

types of extensively modified UNs 'Synonym' while the other rater called them 'User term too fine.' Otherwise there was no significant bias between the raters.

Seventy-one UNs were selected by at least one rater as having medically acceptable concepts that could not be identified in LU although they only agreed on 15. One of the raters submitted these terms and related concepts to the Meta '96 and found that Meta '96 contained close concept matches for 44 of these terms, but 27 terms did not have close concept matches. Nine of the 15 terms identified by both raters could not be found in Meta '96.

DISCUSSION

The Unresolved Narratives Data

The failure to map clinically useful terms to a controlled vocabulary is a multifaceted problem. Areas of potential failure include inadequate coverage of the concepts expressed, inadequate lexical manipulation or synonymy recognition, and user interface issues. In the case of DHCP Problem List, the user interface certainly contributed to the problem by the ease with which it allowed users to generate UNs. In 3-4% of UNs, the user was presented with an exact string match, and in 7-10% of UNs, the user was presented with a good concept match, yet the user opted to generate an UN. Part of the reason for this may be that the closest concept match was not always presented first on the list. This issue is being improved in Version 2 of LU.

The most pervasive problem (both in the case of the LU and our limited test of Metathesaurus) appears to be our limited ability to perform lexical manipulations on the input text string and to identify valid synonyms. 82-84% of the time, the users had experienced negative LU searches, although in only 4-18% was the main concept absent from LU. In the case of Metathesaurus, a Normalized String Index query failed to identify any possible matches in 83% of UNs, yet roughly only 27 UNs (7%) did not have close concept matches. The relatively few terms whose concepts could not be found represent an important potential source for maintenance and updating of both the VA Lexicon Utility and its Metathesaurus base. Submission of subsets of the UNs to the NLM/AHCPR Large Scale Vocabulary Test are planned.

What sort of lexical manipulations would be required to recognize a higher fraction of problems from a patient's problem list? Table 2 presents our breakdown of what these manipulations might be and how frequently they would be required. The data in the 2nd column (selected by both raters) constitutes a lower bound of sorts on the likelihood that the indicated reason contributed to a term not matching. The 3rd column (selected by at least one rater) represents our most liberal estimate of the rate of causation. The most frequent cause of failure to match was the presence of modifiers in 40-53% of UNs. 10% of UNs failed to match because they contained forms of the term "status post", indicating a past procedure history (e.g., "s/p appendicitis"). An additional 1-3% of narratives used a date/time (e.g., "appendicitis 1942"). In 1-3% of narratives, certainty modifiers were applied (e.g., "possible appendicitis"). However, most modifiers causing failure to match were highly diverse and fell into the "other" category (31-47% of UNs). Anatomic structures as in "bibasilar pneumonia," and time course modifiers as in "chronic atrial fibrillation" are examples. In order to adequately deal with this problem, lexical preprocessing routines will almost certainly need to identify and "strip off" these modifiers prior to attempting to match the core concept.

Synonymy was the next most common reason for lack of resolution (13-36% of UNs). In the case of category 2, the user either wanted their exact surface form or they simply made a mistake and failed to pick a good synonym. More often, an alternative term of equal granularity existed within LU, but was not retrieved by the lexical routines utilized for lookup and thus was never presented to the user for consideration. Correction of this problem may require a more extensive synonym table or perhaps performing single word synonym substitutions within a phrase prior to identifying potential concept matches.

Modifiers and synonyms accounted for the bulk of the problems found in the UNs, but the frequency of some of the other reasons was surprising. The use of abbreviations (1-8% of UNs) and misspellings (4-9%) causing failure to match was lower than we had expected. The actual occurrence of these was considerably higher, but MTLU did well in disregarding many misspellings and recognizing many abbreviations. There was a fairly high use of numerical data, such as CPT or ICD codes (11-13%). Although we are unsure of who the exact users were who entered the UNs, we hypothesize that many of the users assigned to Medical Administration services (accounting for 19% of the UNs) were actually clerical staff, and they may have contributed many of the numerical UNs. The category of "not valid" (5-10%) was assigned when the narrative seemed to indicate a non-problem such as "advice or health instruction", which might better have been recorded using other DHCP software. In 6-12% of UNs, the user included what the rater considered more than one concept in a single problem, for example "CVA with Hemiparesis". Concepts were most often related reflecting clinicians' tendency to link pathophysiologically related problems. Constraining physicians to enter one concept per problem will be unlikely to succeed because of this tendency. Strategies for mapping clinician entered terms which may contain more than one concept may require a fundamentally different approach than is currently being used in Problem List/LU and the UMLS/Metathesaurus. The optimum concept match(es) may change depending upon how many concepts are simultaneously sought.

The Inter-Rater Agreement Data and Procedures

The inter-rater agreement data provides some important results for the VA and for other sites who plan to incorporate user feedback in their lexicons. First, the study demonstrated that our raters could be trained, by our Phase I and Phase II procedures, to reliably and reproducibly judge the categorization and causes of the UNs. There was no systematic bias on the part of one of the raters to either "up-rate" or "down-rate" the user-interface response category ratings.

Of the 400 narratives classified, 43% had complete concordance of the raters for each individual cause of failure to match. The inter-rater agreement for the most frequent causes of UNs, modifiers, synonyms, user term too fine, medically acceptable but not in LU, numerical data, and multiple related problems was better than expected by chance. As one might expect based on the manual nature of the task of searching for matching medical concepts, percent agreement on modifiers and numerical data was higher than on synonyms, user term too fine, and concepts not present in the lexicon. There was some bias between the raters on two of the reasons, and some refinement of our definitions of the 'synonym' and 'user term too fine' reasons would be required prior to larger scale study.

CONCLUSION

In this paper, we have developed and validated a reliable process for review of free-text narratives generated in a clinically active VA setting. User interface problems and inadequate lexicon concept coverage explain some of the UNs, but the major problem appears to be poor ability to handle modified terms and synonyms. Using our Phase I and II procedures, we have demonstrated that multiple reviewers can achieve sufficient inter-rater agreement to now proceed with the rating of disjoint narrative sets.

Acknowledgments

The authors would like to thank Dr. Robert Hausam for his input early in the design stage of the study. We would also like to thank Mr. Kimball Rowe who authored the LU. Dr. Hurdle and Dr. Eagon are VA Special Fellows in Medical Informatics. Dr. Eagon is also supported by the Washington University Department of Surgery. Dr. Lincoln is supported by the DVA.

References

1. Cimino JJ. Data storage and knowledge representation for clinical workstations. Int J Biomed Comput. 1994;34:185-94
2. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a controlled medical vocabulary server: The VOSER project. Comput Biomed Res. 1994;27:472-507.
3. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc. 1994;1:142-60.
4. Shiffman S, Detmer WM, Lane CD, Fagan LM. A continuous-speech interface to a decision support system: I. Techniques to accommodate for misrecognized input. J Am Med Inform Assoc. 1995;2:36-45.
5. UMLS Knowledge Sources, 6th Ed. Bethesda, MD: US Dept. of Health and Human Services, NIH, HLM, 1995.
6. Cyr L, Francis K. Measures of clinical agreement for nominal and categorical data: the kappa coefficient. Comput Biol Med. 1992;22:239-46.
7. Lu B. A clinical master dictionary for coding patient problems. Master's Thesis. Salt Lake City, UT: University of Utah, 1995.
8. Siegel S, Castellan, NJ. Nonparametric Statistics for the behavioral sciences. New York, NY:McGraw-Hill, 1988.